Eur J Clin Chem Clin Biochem 1996; 34:145–150

Performance of Immunoassays for CA 19-9, CA 15-3 and CA 125 Tumour Markers Evaluated from an International Quality Assessment Survey

Alessandro Pilo[1], *Gian Carlo Zucchelli*[1], *Richard Cohen*[2], *Maria Rosa Chiesa*[1] and *Charles Albert Bizollon*[2]

[1] CNR, Istituto di Fisiologia Clinica, Pisa, Italy
[2] Service de Radiopharmacie et Radioanalyse, Universitè de Lyon, Lyon, France

Summary: Data collected in the 1993 and 1994 cycles of an international External Quality Assessment (EQA) programme were cumulatively analysed to evaluate the analytical performance of the methods currently in use for routine assay of mucinous tumour markers CA 19-9, CA 15-3 and CA 125. On average the between-laboratory variability was 14.7 and 15.8 CV% for CA 15-3 and CA 125 respectively. For CA 19-9, a markedly worse between-laboratory variability (on average 27.2 CV%) was found; the agreement of CA 19-9 results worsened in the last few years when new non-isotopic techniques became available. The variability component attributable to systematic differences between methods/kits was relatively small for CA 15-3 and CA 125 (17% and 21% of the total variability), while it was markedly larger for CA 19-9 (45% of the total variability). The precision of the methods/kits most often used in the survey ranged from 9.6 to 13.9 CV% for CA 125 and from 10.8 to 14.1 CV% for CA 15-3. For these two tumour markers the precision of the traditional IRMAs does not appear to be different from that of the new fully automated non-isotopic techniques. The precision of CA 19-9 methods was on average worse (from 11.9 to 19.2 CV%), even though the precision of the two automated systems was better than that of IRMAs. In conclusion, the results of this study indicate that the between-laboratory agreement for CA 15-3 and CA 125 assays appears satisfactory, while the CA 19-9 assay shows larger differences between methods and is affected by poorer precision of kits.

Introduction

Determination of mucinous tumour markers CA 19-9, CA 15-3 and CA 125, recognized by monoclonal antibodies, is generally considered a useful tool in the monitoring of cancer patients (1, 2). The assay of these tumour markers is routinely carried out by many laboratories, and numerous methods/kits have been developed and are commercially available. Immunoradiometric (IRMA) techniques with ^{125}I-labelled antibody as a tracer were used earlier for the assay of mucinous markers; more recently non-isotopic immunoassays, based on antibodies labelled with enzymes, fluorescent dyes or chemiluminescent compounds have been developed. The assays performed with these latter techniques can also be carried out with fully automated systems.

The increasing number of different methods/kits available and the large number of samples routinely assayed prompted the setting up of external quality assessment (EQA) schemes to evaluate the analytical performance of the laboratories and of the methods; for this reason the EQA for carcinoembryonic antigen and α-foetoprotein organized by our Institute and sponsored by CNR was extended to the mucinous markers (3, 4). Starting from 1991 the CNR programme joined with the Oncocheck International EQA organized by Service de Radiopharmacie et Radioanalyse, University of Lyon and by Cis BioInternational (5).

Data collected in 1993 and 1994 cycles of Oncocheck EQA have been cumulatively analysed in this paper to evaluate the performance of the routinely used methods.

Materials and Methods

Outline of the Oncocheck EQA program

The Oncocheck programme includes, at present, six tumour markers: α-foetoprotein, carcinoembryonic antigen, CA 19-9, CA 15-3, CA 125 and prostate-specific antigen. The scheme does not substantially differ from other EQAs (6, 7): participants measure 24 samples every year (2 samples each month); they are asked to perform the assay routinely and to return results indicating the method/kit used; collected results are computer processed by the organizing centers of Lyon and Pisa; monthly and cumulative (six month period) reports are prepared and sent back to the participants. At present, the Oncocheck programme involves more than 250 laboratories of many European countries (mainly in Italy and France).

Control materials are normal pools with added sera from patients with high concentrations of tumour markers; control samples (containing all six tumour markers) are prepared from these pools and freeze-dried.

During an EQA cycle (six month) control samples derived from the same pool are mailed out in different batches, as hidden replicates, to estimate the reproducibility of the laboratories and of the kits.

The present analysis is based on results collected from 48 quality control samples distributed during the Oncocheck EQA cycles 1993–1994. The 48 samples were prepared from 24 different pools; 8 pools gave origin to 24 samples (assayed as hidden triplicates), 8 pools gave origin to 16 samples (assayed as hidden duplicates), and 8 pools gave origin to the remaining 8 samples.

Monthly and cumulative reports

Results collected from a control sample are included (as a frequency distribution histogram) in the monthly report, together with the main statistics (mean, median, CV, range) computed from all data and from data grouped according to the method/kit. The aim of this report is to allow comparison of the result produced by a laboratory for a single EQA sample with those produced by all other participants and in particular by the users of the same method/kit.

An end-of-period or cumulative report is also prepared to provide the participant with an estimate of bias and precision based on the results of all control samples assayed during the control cycle. In addition, this cumulative report contains estimates of the analytical performance of the kits most used in the survey; details of monthly and cumulative reports were described previously (8).

Data analysis

The following estimates of variability are used in the evaluation of EQA data and are here briefly recalled; for more details see l. c. (8).

Total variability

The average total variability observed during the whole EQA period (also referred to as between-laboratory agreement or between-laboratory variability) is estimated by averaging the CVs computed from the results of each control sample; this measurement of variability includes both systematic between-kit differences and differences introduced by the laboratories.

Between-kit and within-kit components

A statistical technique (one-factor ANOVA with components of variance estimation (9, 10)) is used to split the total variability into two components: the between-kit variability which accounts for the systematic differences in results produced by different kits, and the within-kit variability which represents the precision of "the average kit".

Kit precision

The precision of the kits is estimated by averaging the CVs of the results produced by participants for the same control material (assayed in different laboratories and on different occasions by the users of the same kit); the reported average CV is therefore an estimate of the between-laboratory, between-assay precision achieved by the kit during the whole EQA period.

Results and Comments

Total variability and within-kit and between-kit components

The total variabilities for CA 15-3, CA 19-9 and CA 125 observed in 48 quality control samples are shown in figure 1 as between-laboratory precision profiles; on average a between-laboratory agreement of 14.7 and 15.8 CV% was observed for CA 15-3 and CA 125 respectively. For CA 125 the variability increases in control samples in the low concentration range (< 17 kIU/l).

The agreement of CA 125 and CA 15-3 assays can be considered satisfactory, since it is similar to or smaller than the between-laboratory variabilities observed for the other three tumour markers included in the Oncocheck EQA (18.6 CV% for carcinoembryonic antigen, 15.3% for α-foetoprotein, 36.0% for prostate-specific antigen).

In contrast, the between-laboratory variability found in CA 19-9 is markedly worse (on average 27.2 CV%); in addition the CVs of CA 19-9 results observed in the 48 control samples differ greatly from each other (ranging from 18.1% to 34.3%) depending on the control material assayed; this behaviour is at variance with that observed for CA 15-3 and CA 125 which showed a narrower range of CVs.

The differences in the variabilities observed in the three tumour markers were further investigated by using the

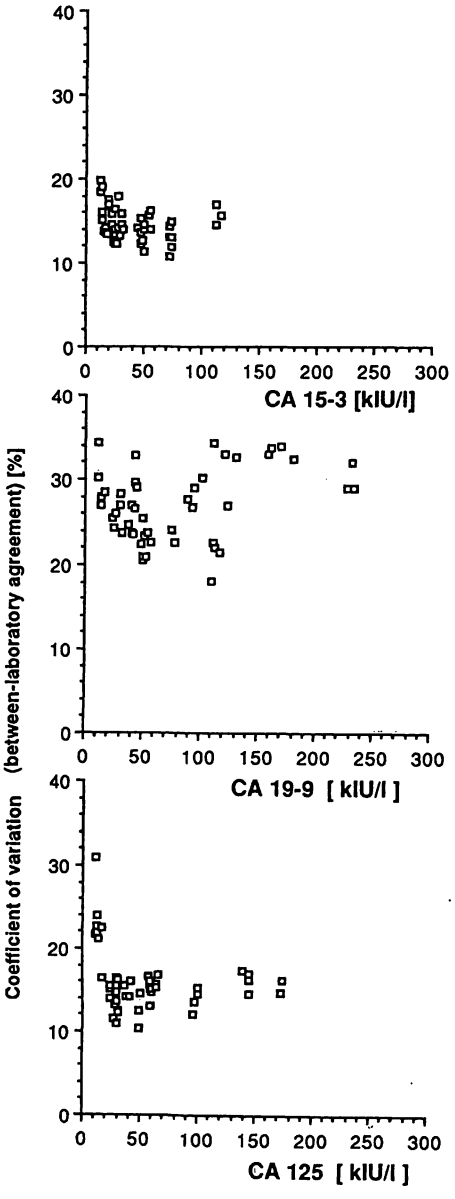

Fig. 1 Total variability (between-laboratory agreement) observed in 48 QC samples plotted against the respective consensus means for CA 15-3, CA 19-9 and CA 125. Data refer to control samples distributed in the Oncocheck programme (from December 1992 to November 1994).

Anova technique to evaluate the between-kit and within-kit components of total variability. The between-kit component (which reflects the systematic differences in results produced by different methods/kits) is relatively small for CA 15-3 and CA 125 (this component accounts only for 17% and 21% of the total variability), while it is markedly larger for CA 19-9 (45% of the total variability). The large between-kit differences in CA 19-9 can be clearly appreciated from figure 2 in which histograms of the results produced by four different methods are reported and compared with the distribution of all data.

The within-kit variability (an estimate of the precision of the "average" kit) was found to be worse in CA 19-9 (20.1 CV%) than in CA 15-3 (13.4 CV%) and CA 125 (14.0 CV%).

Between-kit comparison

CA 15-3

Results collected in the EQA programme were mainly produced by 5 different methods/kits: 3 IRMAs (Centocor, CIS and Byk-Sangtec) and 2 non-isotopic, fully automated techniques (LIA-mat S300 Byk-Sangtec, Enzymun Test ES Boehringer Mannheim). The results produced by these 5 methods on control samples were compared by regression analysis (versus IRMA Centocor) (tab. 1). The results of the 5 different methods appear well correlated (R = 0.98−0.99) and the differences in concentration are on average 10% or less.

CA 125

Results collected in the EQA were mainly produced by 7 different methods/kits: 4 IRMAs (Centocor, CIS, Sorin and Byk-Sangtec) and 3 non-isotopic, fully automated techniques (LIA-mat S300 Byk-Sangtec, Enzymun Test ES Boehringer and IMX Abbott).

The results produced by these 7 methods on control samples were compared by regression analysis (versus IRMA Centocor) (tab. 1). The results from the 7 different methods appear well correlated (R = 0.98−1.00); the differences in concentration are higher than those observed for CA 15-3; for samples with concentrations > 100 kIU/l, three of the methods (IRMA and LIA Byk-Sangtec, Enzymun Test ES Boehringer) give results higher than IRMA Centocor.

CA 19-9

The majority of results collected in the EQA were produced by 8 different method/kits: 4 IRMAs (Centocor, CIS, Sorin and Byk-Sangtec) and 4 non-isotopic, fully automated techniques (LIA-mat S300 Byk-Sangtec which uses a luminescent tracer); and 3 systems (Enzy-

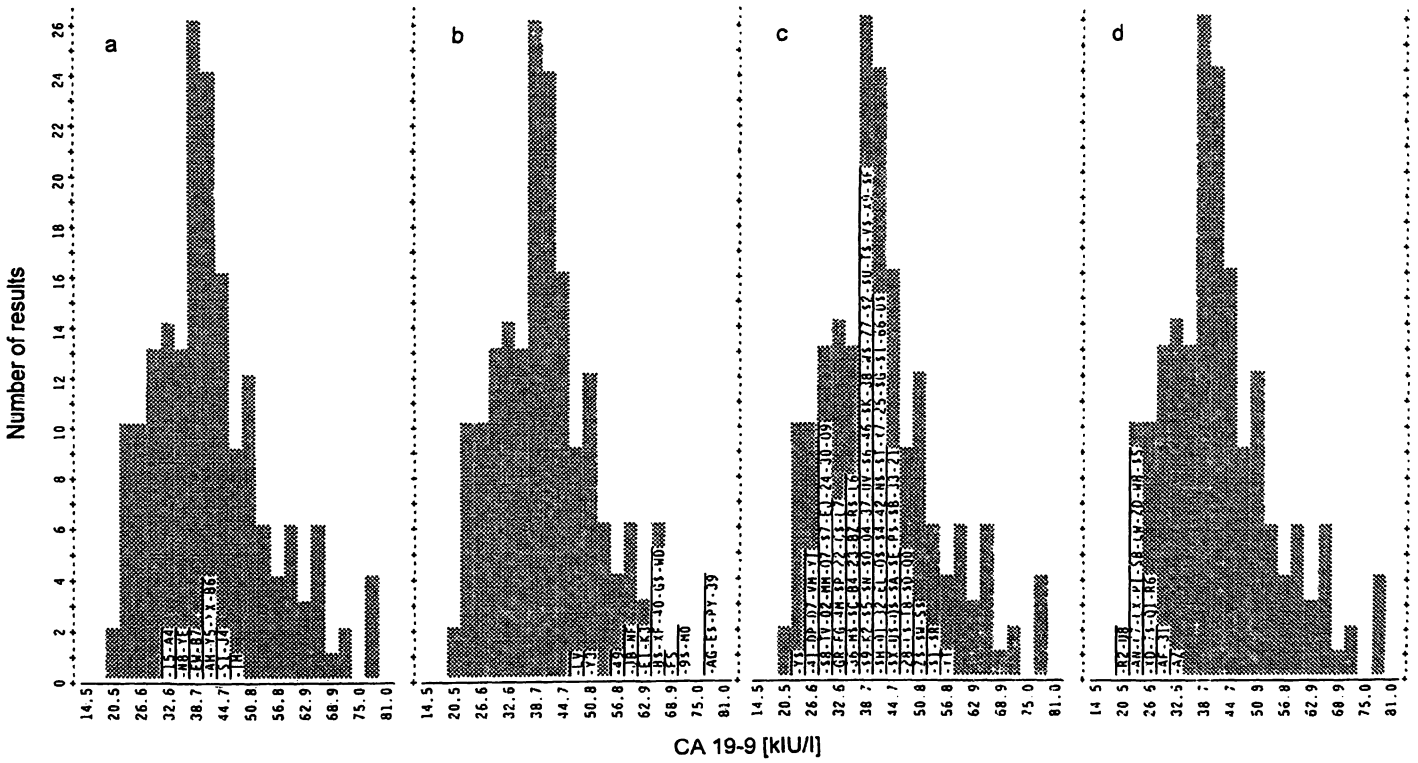

Fig. 2 Frequency distribution histograms of results (identified by the laboratory code) produced by four different methods/kits assaying CA 19-9 in the control sample B1401 distributed in January 1994:
The distribution of all results is represented by the hatched histogram (consensus mean = 43.2 kIU/l, CV = 27.7%, n = 181).

a) IRMA Centocor mean = 41.6 kIU/l, n = 13;
b) IMX Abbott mean = 66.4 kIU/l, n = 17;
c) IRMA Cis mean = 40.1 kIU/l, n = 86;
d) Enzymun Boehringer mean = 26.6 kIU/l, n = 18.

mun Test ES Boehringer, IMX Abbott and Cobas Core Roche, based on the use of enzymes as tracer). Method comparison based on regression analysis is reported in table 2; regression was performed versus IRMA Centocor because this assay was the first routinely used. Results of the 4 IRMAs appear well correlated (R = 0.98–0.99) with relatively small differences in measured concentrations; results from the luminescent method LIA-

Tab. 1 CA 15-3 and CA 125: regression analysis between the method/kits most commonly used in the survey.

x	y	Slope	Intercept (kIU/l)	R
CA 15-3				
IRMA Centocor[a]	IRMA Cis	0.83	+2.8	0.99
IRMA Centocor	LIA Byk Sangtec	1.02	+1.0	0.99
IRMA Centocor	IRMA Byk Sangtec	0.92	+0.4	0.98
IRMA Centocor	Enzymun Test Boehringer	0.87	+1.2	0.98
CA 125				
IRMA Centocor[a]	IRMA Sorin	1.00	+0.4	1.00
IRMA Centocor	IRMA Cis	0.99	−1.0	0.99
IRMA Centocor	IRMA Byk Sangtec	1.34	−9.0	0.99
IRMA Centocor	LIA Byk Sangtec	1.30	−9.9	0.99
IRMA Centocor	IMX Abbott	1.12	−5.3	0.98
IRMA Centocor	Enzymun Test Boehringer	1.20	−6.2	0.98

[a] Regression of the most commonly method/kits vs IRMA Centocor; the concentration range of the samples was 15–138 kIU/l for CA 15-3 and 10–153 kIU/l for Ca 125 (measured by IRMA Centocor).

Tab. 2 CA 19-9: regression analysis between the method/kits most commonly used in the survey.

x	y	Slope	Intercept (kIU/l)	R
IRMA Centocor[a]	IRMA Sorin	0.86	+3.1	0.99
IRMA Centocor	IRMA Byk Sangtec	1.13	−2.0	0.99
IRMA Centocor	IRMA Cis	0.75	+6.1	0.98
IRMA Centocor	LIA Byk Sangtec	1.25	−4.5	0.96
IRMA Centocor	IMX Abbott	1.25	+8.7	0.92
IRMA Centocor	IEMA Roche	1.47	+4.3	0.88
IRMA Centocor	Enzymun Test Boehringer	0.63	+8.8	0.81
IMX Abbott[b]	IEMA Roche	1.21	−9.4	0.98
IMX Abbott	Enzymun Test Boehringer	0.55	−2.0	0.97

[a] regression of the most commonly used method/kits vs. IRMA Centocor; the concentration range of the samples was 10–240 kIU/l (measured by IRMA Centocor).

[b] regression among the three method/kits using enzymes as tracer: IMX Abbott (x), IEMA Roche and Enzymun Test Boehringer (y).

Tab. 3 Comparison of CA 19-9 results (kIU/l) produced by the 8 most popular method/kits[a]).

Pool	IRMA Centocor (12–15)[b])	IRMA Sorin (10–12)	IRMA Byk (12–15)	IRMA Cis (80–90)	LIA Byk (8–10)	IMX Abbott (18–24)	IEMA Roche (8–10)	Enzymun Boehringer (18–22)
P017[c])	17.0	16.3	18.0	14.0	17.8	25.6	26.2	15.9
P030	33.2	30.7	37.4	28.4	38.2	37.8	37.5	18.2
P021	39.4	37.1	37.5	38.0	35.9	77.4	105	54.7
P027	43.8	43.3	49.9	38.5	55.0	64.1	53.3	27.0
P016	73.6	70.5	70.3	70.0	83.5	141	157	89.3
P034	104	98.1	121	113	121	139	129	81.7
P029	106	88.0	119	82.7	135	110	113	53.4
P023	111	102	115	97.6	125	163	205	94.6
P022	208	184	246	153	280	232	280	99.2
P014	232	219	250	202	297	402	488	238

[a]) mean results from 10 pools are reported in this table as examples; the behaviour of the different kits on these pools is representative of that observed in all the control samples circulated in the EQA.

[b]) number of users of the method/kit in the Oncocheck program.

[c]) each pool was sent to the participants on 2–3 occasions as a hidden replicate sample.

mat are also well correlated with IRMAs (R = 0.98). Results of the 3 enzymatic methods are, in contrast, scarcely correlated with IRMAs (R = 0.83−0.93); their results however are well correlated with each other (0.97−0.99 versus IMX Abbott), even though the concentrations found by Enzymun Test are, in all cases, about 50% lower than those found by IMX or Roche.

The poor correlation of enzymatic techniques versus IRMAs can be appreciated in more detail from the mean concentrations found in 10 control pools (representative of all control materials distributed during the EQA) reported in table 3. It can be seen that different results are produced by enzymatic methods and IRMAs, depending on the different control material assayed. This behaviour is exemplified by comparison of 3 pairs of pools shown in figure 3. In pool P030 and in pool P021 the four IRMAs and the luminescent method measure similar concentrations, whereas according to the 3 enzymatic methods the concentration in pool P021 is almost double that of pool P030. Similar behaviour is observed for pool P029 and P023. On the other hand, different behaviour is observed in pools P016 and P034: the three enzymatic methods measure similar concentrations, while IRMAs and LIA find in pool P034 a concentration approximately 30−40% higher than that measured in pool P016.

The reasons for these discrepancies are unclear; it is, however, conceivable that different method/kits for CA 19-9, even though based on the same monoclonal antibody (1116-NS 19-9, from Centocor) and the same antigen for standard preparation (also from Centocor), may show a different degree of specificity against CA 19-9 determinants, due to differences in the tracer, in the solid-phase preparation and/or in the experimental assay conditions (pH, time and temperature of the antigen/antibody reaction). The presence in serum of different molecular forms of CA 19-9 (11), detected by the methods/kits with different degree of specificity, may explain the discrepancies observed in different control samples. Moreover, it is likely that the specificities of IRMAs (and of the chemiluminescent assay) are similar to each other, whereas the three enzymatic techniques also show specificities similar each other, but different from those of IRMAs.

This hypothesis explains why determinations of control samples distributed in the EQA (which probably contain the different CA 19-9 forms in different ratios) correlate only if performed by techniques of the same type (IRMA or IEMA).

Further support for these considerations is provided by data in figure 4, in which results produced by Abbott IMX and Enzymun Test Boehringer (both expressed as fraction of IRMA Centocor) are plotted against the concentrations measured by IRMA Centocor; it can be clearly appreciated that:

1) the two enzymatic methods measure CA 19-9 with specificity different from IRMA and

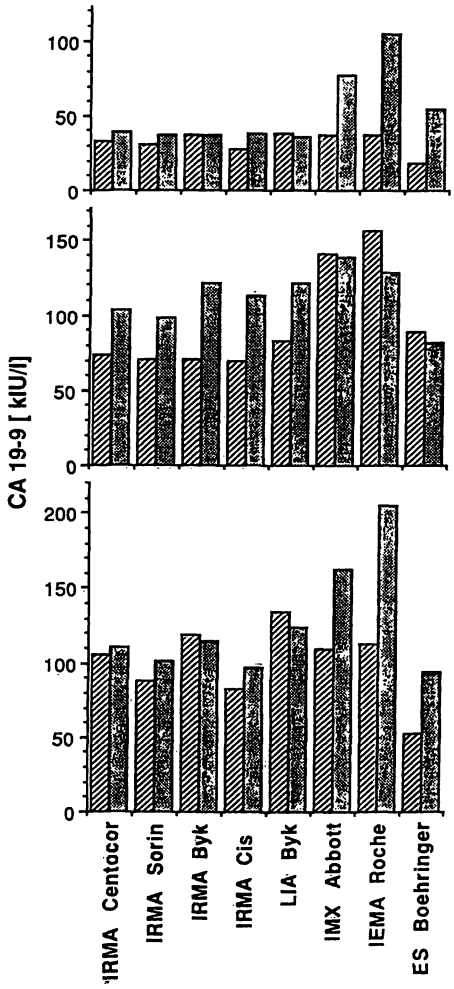

Fig. 3 Mean results produced by the 8 most used methods/kits assaying 6 pools distributed in the EQA period; direct comparisons of 3 pairs of pools (P030 vs P021 top panel; P016 vs P034 middle panel; P029 vs P023 bottom panel) shows the different behaviour of isotopic and enzymatic immunoassays with different control materials.

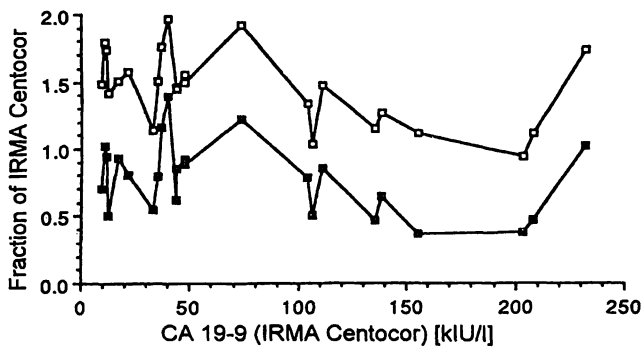

Fig. 4 Mean results found, for the 24 pools distributed during the EQA period (1993−1994), by users of IMX Abbott (□−□) and Enzymun Test Boehringer (■−■), compared with results from IRMA Centocor; data from IMX Abbott and Enzymun Test Boehringer are expressed as fraction of the corresponding results produced by IRMA Centocor.

2) the two enzymatic methods recognize CA 19-9 with the same specificity; the lower values produced by Enzymun Test with respect to IMX can be simply explained by differences in the calibration.

Precision of the method/kits

The precision of the kits was estimated by averaging the CVs of results produced by the kit for the same control material (assayed in different laboratories and on different occasions); therefore the reported average CV can be considered as an estimate of the between-laboratory and between-assay precision achieved by the kit during the whole EQA period.

The precision of the methods/kits most commonly used in the survey is reported in table 4 for the three tumour markers. The CV of the 7 methods for CA 125 assay ranged from 9.6 to 13.9%, and the CV of the 5 methods for CA 15-3 assay ranged from 10.8 to 14.1%. For these two tumour markers the precision of the traditional IRMAs does not appear to differ from that of the new fully automated non-isotopic techniques. On the other hand the precision of the 8 methods used for CA 19-9 assay was found on average to be worse (from 11.9 to 19.2 CV%); it is worth noting that the precision of the automated systems, Abbott IMX and Enzymun Test ES Boehringer, is better than that of IRMAs and similar to that attained in CA 125 and CA 15-3 assays.

Tab. 4 Average precisions[a] (between-laboratory and between-assay, CV%) of the most commonly used methods/kits for CA 15-3, CA 125 and CA 19-9 assays.

Method/kit	CA 15-3	CA 125	CA 19-9
Enzymun Test Boehringer	10.8	11.4	14.2
IRMA Centocor	13.1	10.1	19.2
IRMA Cis	13.2	9.6	18.8
IMX Abbott	–	13.9	11.9
IRMA Byk Sangtec	14.1	11.0	16.8
LIA Byk Sangtec	12.5	12.8	16.1
IEMA Roche	–	–	18.5
IRMA Sorin	–	11.3	18.4

[a] The precisions of the kits were estimated by averaging the CVs of results produced on the same control material assayed in different laboratories and on different occasions. The samples with very low concentration (< 17 kIU/l for the three markers) were excluded from the computation of the average since their CVs were considerably higher than those of all other samples.

Concluding remarks

Analysis of results collected in the EQA for CA 15-3 and CA 125 assays indicates both a satisfactory between-method agreement and a good kit precision. In contrast, the CA 19-9 assay appears to be affected by larger differences between methods and by poorer precision of kits. The agreement of CA 19-9 results worsened in the last few years when new non-isotopic techniques became available and became more widely used for routine purposes (12). These differences may arise because new techniques, in particular those using enzymes as tracer, have different specificities from the IRMA methods.

The satisfactory standardization observed in control samples for CA 15-3 and CA 125 does not however justify over-optimistic conclusions. The possibility cannot be excluded that, due to the heterogeneity of CA 15-3 and CA 125 and to different proportions of their isoforms, method agreement may be worse for patient samples than for control samples.

References

1. Bombardieri E, Seregni E, Crippa F, Buraggi GL. An update on the biology of mucins as tumor markers. In: Up dating on tumor markers in tissues and in biological fluids. Torino: Edizioni Minerva Medica, 1993:477–86.
2. Bon GG, Kenemans P, Van Kamp GJ, Yedema CA, Hilgers J. Review on the clinical value of polymorphic epithelial mucin tumor markers for the management of carcinoma patients. J Nucl Med All Sci 1990; Suppl 34:151–62.
3. Pilo A, Zucchelli GC, Masini S, Torre GC, Ballestra AM. Progress report on an external quality assessment program for immunoassays of tumor markers. J Nucl Med All Sci 1990; Suppl 34:75–82.
4. Zucchelli GC, Pilo A, Chiesa MR, Masini S, Prontera C, Cianetti A. The CNR External Quality Assessment program for tumor markers. In: Up dating on tumor markers in tissues and in biological fluids. Torino: Edizioni Minerva Medica, 1993:129–40.
5. Cohen R, Zucchelli GC, Fraysse M, Pilo A, Rigault MY, Grillet S, Bizollon ChA: "Oncocheck: an international external quality assessment scheme for immunoassays of tumor markers". Nucl Med Biol 1994; 21(3):483–93.
6. Seth J, Sturgeon CM, Al Sadie R, Hanning I, Ellis AR. External quality assessment of immunoassay of peptide hormones and tumour markers: principles and practice. Ann Ist Super Sanità 1991; 27:443–52.
7. Cohen R, Bizollon ChA. Immunoassay external quality assessment schemes in France. Ann Ist Super Sanità 1991; 27:503–10.
8. Pilo A, Zucchelli GC, Chiesa MR, Masini S, Clerico A. The CNR external quality assessment program for immunoassays: statistical analysis and reports for participants. Ann Ist Super Sanità 1991; 27:469–78.
9. McDonough F, Munson PJ, Rodbard DA. A computerized approach to statistical quality control for RIA in clinical chemistry laboratory. Comput Prog Biomed 1977; 7:179–85.
10. Pilo A, Zucchelli GC, Chiesa MR, Bolelli GF, Albertini A. Components of variance analysis of data produced in a national quality control survey of radioimmunoassay of T3, T4, TSH, prolactin and progesterone. Clin Chem 1986; 32:171–5.
11. Hammarstroem S. Chemistry and immunology of CEA, CA 19-9 and CA 50. In: Holmgren J, editor. Tumor markers antigens. Studentlitteratur, Lund 1985:34–51.
12. Zucchelli GC, Pilo A, Chiesa MR, Cohen R, Bizollon CA. Growing use of nonisotopic CA 19-9 immunoassays increases between-laboratory variability. Clin Chem 1993; 39:909–11.

Received June 12/November 3, 1995

Corresponding author: dr. A. Pilo, CNR, Istituto di Fisiologia Clinica, via Savi 8, I-56100 Pisa, Italy

Eur J Clin Chem Clin Biochem 1996; 34:151–158

Evaluation of the Abbott IMx Ultrasensitive II hTSH Immunometric Assay in Three European Centres: A Comparison with Established Commercial Immunometric Assays for Thyrotropin

William Graham Wood[1], *Ulrike Bruns*[1], *Otto Eber*[2], *Werner Langsteger*[2], *Jean Yves Bounaud*[3] and *Marie-Paule Bounaud*[3]

[1] Institut für Klinische Laboratoriumsdiagnostik, Klinikum der Hansestadt Stralsund, Stralsund, Germany
[2] Interne Abteilungen des Krankenhaus der Barmherzigen Brüder, Graz-Eggenberg, Austria
[3] Service Biophysique, Hôpital Jean-Bernard, CHU, Poitiers, France

Summary: The Abbott Ultrasensitive II hTSH assay was evaluated in three European centres and its performance compared with its predecessor and with commercially available kits. A total of 408 individuals was included in the study (140 euthyroid, 116 hyperthyroid and 86 hypothyroid subjects, as well as 26 patients with non-thyroidal illness and 30 patients with thyroid cancer). The kit was evaluated for (im)precision and analytical and functional sensitivity according to ECCLS-Guidelines.

The analytical sensitivity lay between 0.004 and 0.013 mU/l, the mean value being 0.008 mU/l, results being from 12 runs. The functional sensitivity gave a coefficient of variation below 20% at a concentration of 0.05 mU/l under routine conditions.

Inter-assay precision was less than 7% at 0.25 mU/l (range 5.3–6.8%), less than 6% at 6 mU/l (range 4.0–5.2%) and less than 9% at 30 mU/l (range 6.5–8.7%). Intra-assay (im)precision was not calculated as the Abbott IMx is designed to run on singlicate determinations.

Introduction

The aim of the study was to evaluate the new Ultrasensitive II hTSH assay designed for use on the Abbott IMx by comparing performance with its predecessor and with three well established commercial immunometric assays designed to measure thyrotropin in human serum/plasma samples.

The study was carried out in laboratories in Germany, France and Austria using an Abbott IMx coupled to a computer for direct data reduction and storage on floppy disks.

The main points examined were the (im)precision (according to ECCLS format), the analytical sensitivity and the functional sensitivity of the assay, the latter using a panel of seven sera with low thyrotropin concentrations common to all centres.

Clinically relevant samples were chosen to compare the concordance between diagnosis and measured thyrotropin concentration in treated and untreated hypo- and hyperthyroid patients, as well as in patients with no evidence of thyroid disease who acted as euthyroid controls.

Materials and Methods

Kit under examination

The kit under examination was the IMx Ultrasensitive hTSH II, List number 4B01, Abbott Laboratories, Diagnostics Division, Abbott Park, IL, USA. The test was based on an automated microparticle immunoenzymometric assay with alkaline phosphatase as label and 4-methylumbelliferyl phosphate as substrate. Several lots of reagents were used in the study, the results, however, being used without specific reference to the reagent lot, as one aim of the study was to evaluate under normal laboratory conditions.

Kits used for comparison

BeriLux hTSH – Product No. OCNA, Behringwerke AG, Marburg a. d. L., Germany. This kit is an immunoluminometric assay based on coated tube technology and an acridinium derivative as label, and uses a two point calibration against a lot specific master curve (0–100 mU/l thyrotropin), the latter being given into the instrument before using the reagent lot. The euthyroid range was declared by the manufacturer as 0.25–4 mU/l, hyperthyroid patients being "preponderantly less than 0.1 mU/l" and hypothyroid patients having concentrations above 5 mU/l. These ranges were evaluated from 451 individuals.

DYNOtest TSH, B. R. A. H. M. S. Diagnostica GmbH (formerly Henning-Berlin), Berlin, Germany. This immunoradiometric assay is also based on coated tube technology and uses a ^{125}Iodine-label as tracer. A standard curve consisting of seven points (0.02–80 mU/l thyrotropin) is set up with each run. The ranges of concentrations given by the manufacturer for different clinical situations were 0.3–4 mU/l for healthy individuals with no evidence of thyroid disease, euthyroidism < 4 mU/l, preclinical hyperthyroidism < 0.5 mU/l, hyperthyroidism < 0.2 mU/l (mostly < 0.1 mU/l), hypothyroidism > 4 mU/l.

Magic Lite TSH hs, CIBA-Corning GmbH, Fernwald, Germany. This test is an immunoluminometric assay using an acridinium compound as label and magnetic microparticles as solid phase. The calibration principle is identical with that of the BeriLux test and

uses a lot specific master curve with two point calibration for each run. The euthyroid range was declared as 0.25–4.5 mU/l, determined from 241 clinically euthyroid subjects. The euthyroid range established in the laboratory was 0.2–3.5 mU/l.

Ultrasensitive TSH, Abbott Diagnostica, Wiesbaden-Delkenheim, Germany. The method is automated for the Abbott IMx and is a microparticle based immunoenzymometric assay with alkaline phosphatase and 4-methylumbelliferyl phosphate as substrate. The detection is based on reflected fluorimetric. Samples from 405 apparently healthy individuals gave rise to a euthyroid range from 0.32–5 mU/l thyrotropin (2.5–97.5 percentiles). Hyperthyroid patients (n = 39) gave values < 0.2 mU/l and 32 hypothyroid subjects had values above 8 mU/l.

Patient groups studied

A total of 408 individuals was included in the study. These were divided into the following groups according to clinical diagnosis: 140 euthyroid, 116 hyperthyroid (including 20 with subclinical hyperthyroidism [blunted thyroliberin-response]), 86 hypothyroid, 26 non-thyroidal illness and 30 with thyroid cancer.

Tab. 1a Functional sensitivity data from all centres for both the test and comparison methods

German centre	Abbott Ultra II		BeriLux	
Panel member	Mean thyrotropin (mU/l)	CV (%)	Mean thyrotropin (mU/l)	CV (%)
Serum 1	0.010	28.8	0.022	28.0
Serum 2	0.020	21.0	0.039	18.0
Serum 3	0.040	15.1	0.067	18.2
Serum 4	0.050	9.11	0.091	16.0
Serum 5	0.080	8.11	0.119	11.4
Serum 6	0.110	6.50	0.164	9.88
Serum 7	0.150	4.59	0.214	11.7

Austrian centre	Abbott Ultra II		Dynotest	
Panel member	Mean thyrotropin (mU/l)	CV (%)	Mean thyrotropin (mU/l)	CV (%)
Serum 1	0.010	65.7	0.033	34.7
Serum 2	0.020	23.6	0.052	36.3
Serum 3	0.040	18.0	0.108	44.8
Serum 4	0.060	16.9	0.118	21.3
Serum 5	0.080	8.13	0.153	13.4
Serum 6	0.110	11.0	0.203	11.4
Serum 7	0.115	6.67	0.279	10.3

French centre	Abbott Ultra II		Abbott Ultra*	
Panel member	Mean thyrotropin (mU/l)	CV (%)	Mean thyrotropin (mU/l)	CV (%)
Serum 1	0.020	78.3	0.020	79.6
Serum 2	0.030	22.6	0.040	76.6
Serum 3	0.050	15.1	0.070	49.4
Serum 4	0.070	7.82	0.090	32.5
Serum 5	0.090	11.7	0.120	19.5
Serum 6	0.120	9.27	0.150	18.6
Serum 7	0.150	4.99	0.210	16.5

Key: * The French site compared the new (Ultra II) TSH-Test with its predecessor (Ultra) in this experiment. Data from the test kit (CIBA-Corning Magic-Lite) used for the rest of the study were not available (see l. c. (4) for characteristics of this kit).

The thyroid cancer patients were under thyroxine therapy (100–190 µg/d – mean 150 µg/d) (Euthyrox, Merck) or a combination of 150 µg thyroxine and 20 µg triiodothyronine per day (Novothyral, Merck).

A euthyroid status was assumed when the sonographic picture was normal, the thyroid analytes thyrotropin and free thyroxine lay within the euthyroid reference range and the patient had no symptoms of a thyroid disorder.

Procedures used

Precision

Between-run-precision was determined according to the ECCLS format using three IMx Ultrasensitive hTSH II assay controls and three human based serum panels. Samples were tested in duplicate, once a day on each of ten days. Within-run-precision was performed on 20 replicates of each of three assay controls.

Tab. 1b Summary of between-run precision for the test and comparison methods

German centre	Abbott Ultra II		BeriLux	
Control Serum	Mean thyrotropin (mU/l)	CV (%)	Mean thyrotropin (mU/l)	CV (%)
Low control	0.25	6.8	0.44	7.4
Medium control	6.02	5.2	9.56	11.6
High control**	28.6	6.5	48.9	17.3
Assay panel 1	0.07	11.1	0.08	17.5
Assay penel 2	1.47	5.8	1.89	11.6
Assay panel 3	30.5	8.7	30.8	14.6

Austrian centre	Abbott Ultra II		Dynotest	
Control serum	Mean thyrotropin (mU/l)	CV (%)	Mean thyrotropin (mU/l)	CV (%)
Low control	0.24	5.3	0.4	7.1
Medium control	5.96	5.7	9.05	5.6
High control**	27.1	12.6	45.1	4.6
Assay panel 1	0.05	15.5	***	***
Assay panel 2	1.40	5.3	1.56	9.4
Assay panel 3	31.7	6.5	23.3	17.0

French centre	Abbott Ultra II		Abbott Ultra*	
Control serum	Mean thyrotropin (mU/l)	CV (%)	Mean thyrotropin (mU/l)	CV (%)
Low control	0.27	5.9	0.26	13.0
Medium control	6.29	4.0	6.57	4.0
High Control**	28.6	6.6	56.3	3.5
Assay panel 1	0.08	8.5	0.10	32.8
Assay panel 2	1.45	3.6	1.69	5.0
Assay panel 3	28.9	6.3	29.5	4.4

Key: * The French site compared the new (Ultra II) TSH-Test with its predecessor (Ultra) in this experiment. Data from the test kit (CIBA-Corning Magic-Lite) used for the rest of the study were not available (see l. c. (4) for characteristics of this kit).
** The nominal concentration of this serum was 30 mU/l thyrotropin for the Ultra II test. A separate serum with a nominal value of 50 mU/l thyrotropin was used for the comparison tests.
*** No results were obtained for this serum due to wrong labels on the test serum bottles.
Apart from the high control, identical materials were used for test and comparison assays.

Analytical sensitivity

The potential lower detection limit of the assay was estimated as follows: The zero calibrator was set up tenfold, the remaining five calibrators in duplicate and the standard curve plotted from the results. The potential lower limit of detection was calculated as the concentration on the standard curve equivalent to the mean zero-calibrator rate plus two standard deviations of the zero-calibrator rate. A total of twelve independent runs was performed to determine the analytical sensitivity.

Functional sensitivity

Seven panel members were prepared from human serum containing very low concentrations of thyrotropin. Each of these panels was run in singlicate – i.e. under routine conditions – on each of ten

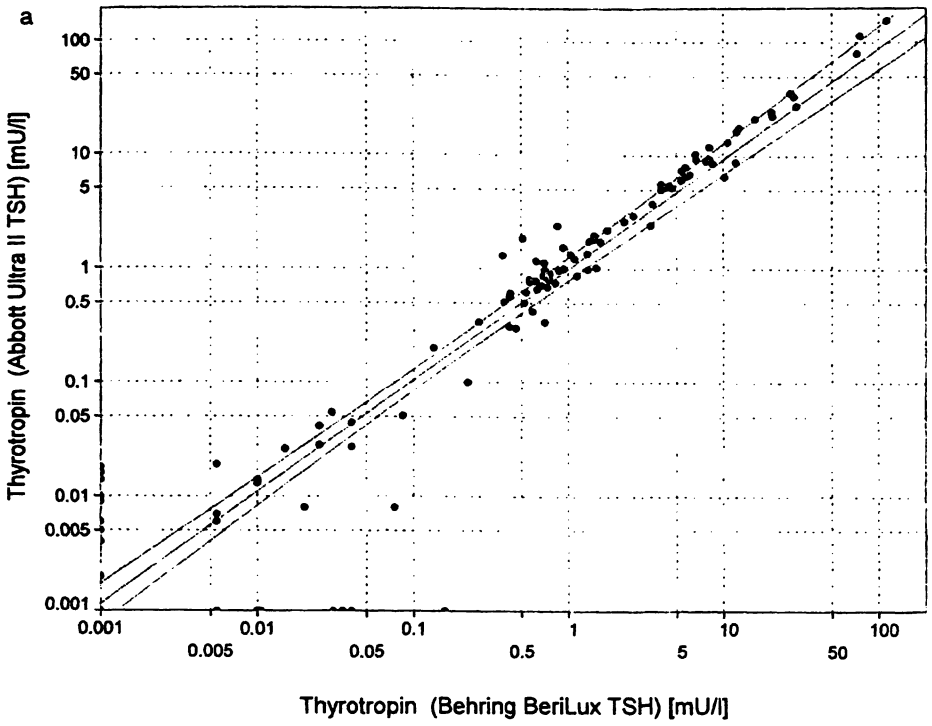

Fig. 1a Data from the German centre.
The data from all samples are plotted on the graph, using a double logarithmic scale because of the skewed distribution of the measured concentrations. This presentation also highlights the samples with low thyrotropin concentrations. The *Spearman* rank correlation coefficient was 0.948 for the 108 samples used. The house-internal kit was plotted on the abscissa, the kit under test on the ordinate. The equation for the regression line was log(y) = 0.997 log(x).

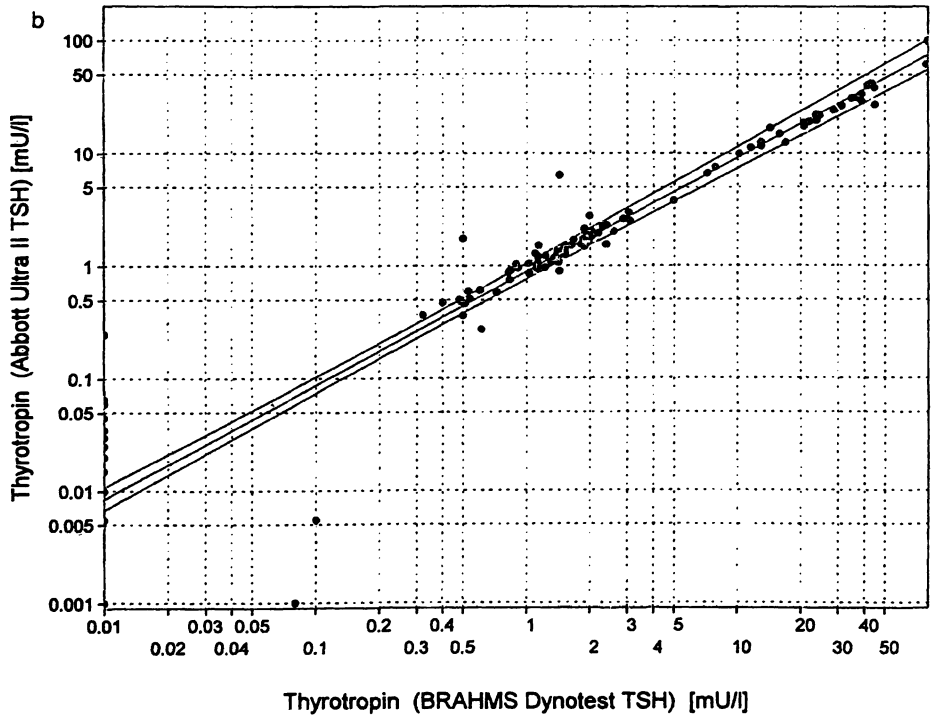

Fig. 1b Data from the Austrian centre.
The data were presented as in figure 1a. The *Spearman* rank correlation coefficient was 0.978 for 106 samples used. The regression line equation was log(y) = 0.981 log(x) + 0.009.

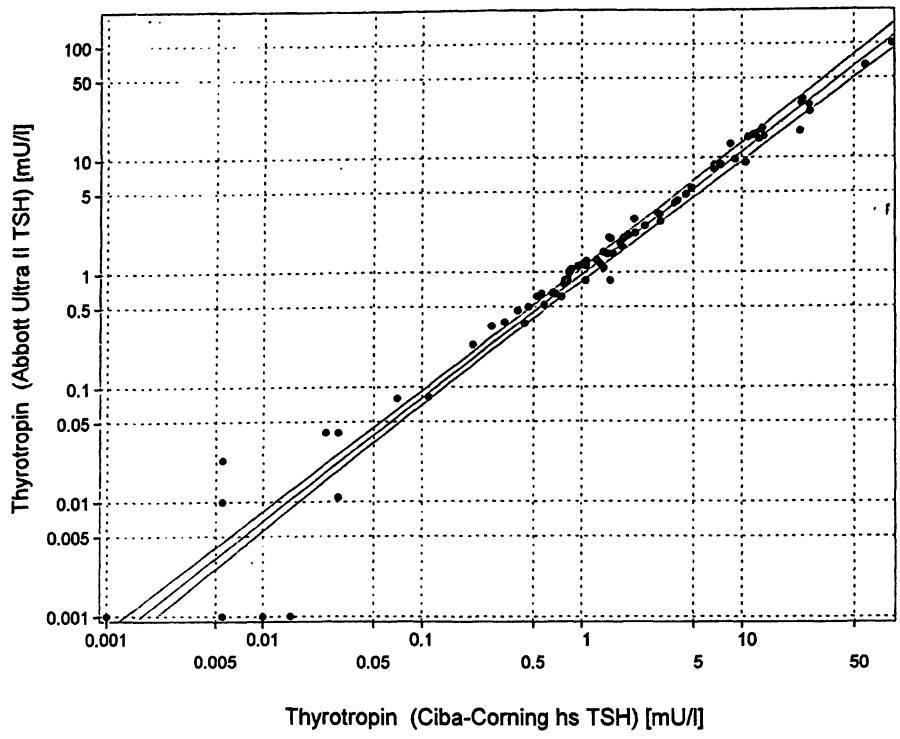

Fig. 1c Data from the French centre.
The data were presented as in figure 1a. The *Spearman* rank correlation was 0.989 and the equation for the regression line: log(y) = 0.987 log(x) − 0.002 for the 96 cases.

Tab. 2a Statistical data from the German centre for all three groups of patients

Euthyroid				
Regression data	n = 48	a = 0.051	b = 0.810	r = 0.825
Median age:	41 years	Mean age:	40 years	
	Ultra-II TSH (x)	Berilux TSH (y)		
Mean (mU/l)	1.01	1.18		
Median (mU/l)	0.74	0.94		
Minimum value	0.14	0.20		
Maximum value	3.51	3.78		
2.5th percentile	0.20	0.24		
97.5th percentile	3.40	3.62		
Hyperthyroid				
Regression data	n = 49	Calculation of a, b & r not done		
Median age:	57 years	Mean age: 54 years		
	Ultra-II-TSH (x)	Berilux TSH (y)		
Mean (mU/l)	0.020	0.012		
Median (mU/l)	Not detectable	Not detectable		
Minimum value	Not detectable	Not detectable		
Maximum value	0.225	0.100		
95th percentile	0.085	0.051		
Not detectable	36/49	21/49		
Hypothyroid				
Regression data	n = 30	a = 1.86	b = 0.717	r - 0.988
Median age:	57 years	Mean age:	54 years	
	Ultra-II TSH (x)	Berilux TSH (y)		
Mean (mU/l)	18.6	23.3		
Median (mU/l)	8.35	9.37		
Minimum value	4.00	5.06		
Maximum value	112	156		
5th percentile	4.10	5.27		

Key: All concentrations are in mU/l.
The number of samples which were below the detection limit in each assay in the hyperthyroid patients is given in the last row of the data for this group. No regression line was constructed for this group as too many values lay below the assay detection limit. The relevant percentiles are given to allow an estimate for any overlap between groups.

days. The mean and coefficient of variation for each panel serum were calculated. The functional sensitivity was recorded visually on precision profile plots (see Results for details).

Recovery and linearity

Recovery and linearity were checked in one laboratory using additive and dilution techniques.

Statistics

Parametric statistics were used in all cases except the regression analysis of all detectable samples (310/408 samples), where a double logarithmic transformation of data was performed before performing the *Spearman* rank correlation between in-house test and the IMx Ultrasensitive hTSH II assay.

Results

Analytical sensitivity

The results from the 12 runs (see above) gave results ranging from 0.004 and 0.013 mU/l, the mean value being 0.008 mU/l. In comparison, the corresponding mean value from the Abbott Ultrasensitive hTSH (the predecessor to the assay under evaluation) was 0.026 mU/l (range 0.019–0.036).

Functional sensitivity

Table 1a shows the results for the seven test panel sera for both the Abbott Ultrasensitive hTSH II and for the kit used for comparison. The results show that the kit under evaluation fits the definition of a 'third generation' thyrotropin assay according to *Nicoloff & Spencer* (1). Only one of the kits used for comparison fulfils this definition (BeriLux). The results must be treated with care, as the evaluation of the functional sensitivity was performed on pooled sera and not from single sera from hyperthyroid/thyrotoxic patients.

Precision

Table 1b summarises the inter-assay (im)precision data. The kit under test showed better precision on the whole, when compared with the test kits used for comparison. This may very well be a result of the

Tab. 2b Statistical data from the Austrian centre for all three groups of patients

Euthyroid				
Regression data	n = 48	a = 0.089	b = 0.881	r = 0.927
Median age:	59 years	Mean age:	59 years	
	Ultra-II TSH (x)	Dynotest TSH (y)		
Mean (mU/l)	1.49	1.59		
Median (mU/l)	1.48	1.52		
Minimum value	0.09	0.11		
Maximum value	2.98	3.08		
2.5^{th} percentile	0.59	0.60		
97.5^{th} percentile	2.78	3.04		
Hyperthyroid				
Regression data	n = 39	Calculation of a, b, & r not done		
Median age:	37 years	Mean age: 40 years		
	Ultra-II TSH (x)	Dynotest TSH (y)		
Mean (mU/l)	0.020	Not detectable		
Median (mU/l)	0.020	Not detectable		
Minimum value	Not detectable	Not detectable		
Maximum value	0.250	Not detectable		
95^{th} percentile	0.060	Not detectable		
Not detectable	23/39	39/39		
Hypothyroid				
Regression data	n = 32	a = −4.25	b = 1.08	r = 0.952
Median age:	44 years	Mean age:	41 years	
	Ultra-II TSH (x)	Dynotest TSH (y)		
Mean (mU/l)	32.0	33.6		
Median (mU/l)	23.2	26.6		
Minimum value	3.80	4.97		
Maximum value	100	81.3		
5^{th} percentile	7.20	7.50		

Key: All concentrations are in mU/l.
The number of samples which were below the detection limit in each assay in the hyperthyroid patients is given in the last row of the data for this group.
The relevant percentiles are given to allow an estimate for any overlap between groups.
The correlation data for the hyperthyroid patients could not be calculated as all data from the Dynotest lay below the detection limit of the assay (0.03 mU/l).

full automation combined with the quality of the antibodies used.

Recovery and linearity

Thyrotropin was added to three sera with endogenous thyrotropin concentrations of 1.62, 2.38 and 4.36 mU/l, the amounts added being 20, 40, 60 and 80 mU/l. The recovery (expected value) lay between 91 and 111% (mean value 100%).

Two sera, one with 0.86 mU/l and one with 49.2 mU/l were diluted with a serum with a thyrotropin concentration under 0.01 mU/l in five steps between 1 : 1.125 and 1 : 10. The values found ranged between 89 and 117% (mean value 104%) of the expected values.

Assay comparison

The assays were compared in different concentration ranges, i.e. with hyper-, eu- and hypothyroid patient sera. In addition, patients with thyroid cancer and non-thyroidal illness were investigated at one test centre. The results were interpreted both in terms of analytical and clinical performance, in order to estimate misclassification errors and where possible the source of error. The in-house kit results are plotted on the abscissa, the kit under test on the ordinate.

Figures 1a–1c show the comparison between the kits used, using a double logarithmic scale for the data sets in each case. Tables 2a–2c show the corresponding data for euthyroid, hyper- and hypothyroid subjects. Table 2d shows the results for the thyroid cancer patients and those with non-thyroidal illness.

Results which gave rise to clinically discordant diagnoses were only found in two cases. One patient classified as having a "non-thyroidal illness" had a thyrotropin concentration in the Abbott test of 6.4 mU/l, in the Dynotest 1.4 mU/l. The corresponding free thyroxine and free triiodothyronine concentrations were 5.2 pmol/l and 2.0 pmol/l, respectively and lay in the hypothyroid range. The second case, a patient with "subclinical hy-

Tab. 2c Statistical data from the French centre for all three groups of patients

Euthyroid				
Regression data	n = 44	a = 0.078	b = 0.895	r = 0.961
Median age:	42 years	Mean age:	45 years	
	Ultra-II TSH (x)	Corning TSH (y)		
Mean (mU/l)	1.21	1.26		
Median (mU/l)	1.04	1.11		
Minimum value	0.07	0.08		
Maximum value	3.11	3.30		
2.5th percentile	0.21	0.24		
97.5% percentile	3.08	3.20		
Hyperthyroid				
Regression data	n = 28	Calculation of a, b & r not done		
Median age:	55 years	Mean age: 55 years		
	Ultra-II TSH (x)	Corning TSH (y)		
Mean (mU/l)	0.006	Not detectable		
Median (mU/l)	0.020	Not detectable		
Minimum value	Not detectable	Not detectable		
Maximum value	Not detectable	Not detectable		
95th percentile	0.035	0.057		
Not detectable	24/28	23/28		
Hypothyroid				
Regression data	n = 32	a = −4.24	b = 1.08	r = 0.952
Median age:	44 years	Mean age:	41 years	
	Ultra-II TSH (x)	Corning TSH (y)		
Mean (mU/l)	17.4	19.9		
Median (mU/l)	11.4	14.8		
Minimum value	3.82	4.03		
Maximum value	86.7	100		
5th percentile	4.10	4.30		

Key: All concentrations are in mU/l.
The number of samples which were below the detection limit in each assay in the hyperthyroid patients is given in the last row of the data for this group.
The relevant percentiles are given to allow an estimate for any overlap between groups.
The correlation data for the hyperthyroid patients could not be calculated as the majority of data from both methods by below the detection limit of the assay.

perthyroidism" had a thyrotropin concentration of 0.25 mU/l in the Abbott test and < 0.03 mU/l in the Dynotest, the corresponding free thyroxine and free triiodothyronine concentrations being 23.4 pmol/l and 5.2 pmol/l, respectively, i.e. both in the euthyroid range. Full concordance of results, as seen from the clinical diagnostic point of view, was achieved in all other patients with defined thyroid function.

A point to note was the relatively poor correlation between both kits tested for the non-thyroidal illness patients. These patients were often undergoing intensive care with treatment involving many drugs and infusions, which may account for the poor comparability of results.

Discussion

The need for thyrotropin assays with low detection limits and good analytical and diagnostic sensitivity has led to continual methodological improvements and the introduction of such terms as 'first', 'second' and 'third' generation assays, depending upon the lower detection limit and (im)precision data of the tests in question (1–4). The aim of the producer of kits to measure thyrotropin must be to bring a kit onto the market which fulfils the actual clinical and analytical needs (5, 6). In the case of thyroid illness, it has been difficult to develop methods able to discriminate fully between eu- and hyperthyroidism, especially using an enzyme label. By combining a microparticle technology with fluorescence enhancement and a fully automated test procedure, (excluding the pipetting of sample), it has become possible to reach the lower detection limit reserved for luminescence, time-resolved fluorescence and radioisotopic labels. In addition, the fluorescence detection overcomes the typical limits set by the *Beer-Lambert* law with colorimetric measurements using a single wavelength for detection.

Although the definition of the lower detection limit or analytical sensitivity of an immunoassay was published many years ago by *Rodbard* (7) and *Ekins* (8), there has been much discussion about this point which has led to a certain amount of confusion. Although the method used in this study for defining the lower detection limit of the assay differs from that of *Rodbard*, it is equally

Tab. 2d Statistical data from the Austrian centre for the sub-clinical hyperthyroid, non-thyroidal illness and thyroid cancer patients

Non-thyroidal illness				
Regression data	n = 17	a = 0.166	b = 1.077	r = 0.461
Median age:	Not given	Mean age:	Not given	
	Ultra-II TSH (x)	Dynotest TSH (y)		
Mean (mU/l)	1.10	0.86		
Median (mU/l)	0.60	0.84		
Minimum value	0.01	0.10		
Maximum value	6.40	2.38		
Clinically sub-hyperthyroid				
Regression data	n = 20	Calculation of a, b & r not done		
Median age:	40 years	Mean age: 43 years		
	Ultra-II TSH (x)	Dynotest TSH (y)		
Mean (mU/l)	0.020	Not detectable		
Median (mU/l)	0.008	Not detectable		
Minimum value	Not detectable	Not detectable		
Maximum value	0.245	0.080		
Non-detectable	12/17	16/17		
Thyroid cancer				
Regression data	n = 30	Calculation of a, b & r not done		
Median age:	40 years	Mean age: 43 years		
	Ultra-II TSH (x)	Dynotest TSH (y)		
Mean (mU/l)	Not detectable	Not detectable		
Median (mU/l)	Not detectable	Not detectable		
Minimum value	Not detectable	Not detectable		
Maximum value	0.060	0.060		
Non-detectable	19/30	26/30		

Key: All concentrations are in mU/l.
The number of samples which were below the detection limit in each assay in the hyperthyroid patients is given in the last row of the data for this group.
The relevant percentiles are given to allow an estimate for any overlap between groups.
The correlation data for the hyperthyroid patients could not be calculated as all data from the Dynotest lay below the detection limit of the assay (0.03 mU/l).

valid statistically. The actual lower detection limit, as defined by the functional sensitivity is a practical rather than a theoretical approach to the problem, although different authors give different performance criteria for the estimation of the functional sensitivity (9–11).

Tables 2a–2d also show the median ages of the groups under study in the three centres. It can be seen that the euthyroid patients in the Austrian study were significantly older than those in the other centres. The mean and median thyrotropin concentrations were higher than in the other centres and were more normally distributed (mean/median ≈ 1.0). The hyperthyroid patients were younger in the Austrian centre than in the other two centres. The hypothyroid patients were older in the German centre than in either of the other two centres.

The good correlation data in the eu- and hypothyroid ranges shows the comparability with other commercial available tests for thyrotropin. The relatively poor correlation in the non-thyroidal illness group (tab. 2d) may come from the limited concentration range of thyrotropin in this group as well as the low number of samples tested (n = 17). The clinical discrimination of samples was excellent, with only 2 aberrant values (see above).

Both the analytical and functional sensitivity of the Abbott Ultrasensitive hTSH II assay place it in the category of a 'third generation' test (1), thus allowing for discrimination between eu- and hyperthyroid patients at the highest level of commercially available analytical testing at the present time.

References

1. Nicoloff JT, Spencer CA. The use and misuse of sensitive thyrotropin assays. J Clin Endocrinol Metab 1990; 71:553–8.
2. Marschner I, van Thiel D, Wood WG, Scriba PC. Vergleich dreier Ringversuche zur radioimmunologischen Thyrotropin-Bestimmung nach dem ‚Münchner Modell'. J Clin Chem Clin Biochem 1983; 21:789–97.
3. Wood WG, Waller D, Hantke U. An evaluation of six solid-phase thyrotropin (TSH) kits. J Clin Chem Clin Biochem 1985; 23:461–71.
4. Bounaud MP, Piketty ML, Breton I, Bounaud JY, Begon F, Marechaud R. TSH de ‹ troisième génération › (TSH hs – Ciba Corning): caractéristiques analytiques, valeur diagnostique pour l'hyperthyroïde. Immunoanal Biol Spéc 1992; 34:45–8.
5. Spencer CA, Schwarzbein D, Guttler RB, LoPresti JS, Nicoloff JT. Thyrotropin (TSH)-releasing hormone stimulation test response employing third and fourth generation TSH assays. J Clin Enocrinol Metab 1993; 76:494–8.
6. Klee GG, Hay ID. Role of thyrotropin measurements in the diagnosis and management of thyroid disease. Clin Lab Med 1993; 13:673–82.
7. Rodbard D. Statistical estimation of the minimal detectable concentration ('sensitivity') for radioligand assays. Anal Biochem 1978; 90:1–12.
8. Ekins R. Assay design and quality control. In: Bizollon CA, editor. Radioimmunology. Amsterdam, New York, Oxford: Elsevier/North Holland Biomedical Press, 1979:239–55.
9. Woodhead JS, Weeks I. Circulating thyrotropin as an index of thyroid function. Ann Clin Biochem 1985; 22:455–9.
10. Bayer MF. Performance criteria for appropriate characterization of (highly) 'sensitive' thyrotropin assays. Clin Chem 1987; 33:630–1.
11. Spencer CA. Thyroid profiling for the 1990's: fT_4 estimate or sensitive TSH measurement. J Clin Immunoassay 1989; 12:82–9.

Received August 9/November 3, 1995

Corresponding author: Prof. Dr. W. G. Wood, Institut für Klinische Laboratoriumsdiagnostik, Klinik der Hansestadt Stralsund, Postfach 2341. D-18410 Stralsund, Germany

Eur J Clin Chem Clin Biochem 1996; 34:159–165 © 1996 by Walter de Gruyter · Berlin · New York

External Quality Assessment: Currently Used Criteria for Evaluating Performance in European Countries, and Criteria for Future Harmonization

Carmen Ricós[1], *Henk Baadenhuijsen*[2], *Jean-Claude Libeer*[3], *Per Hyltoft Petersen*[4], *Dietmar Stöckl*[5], *Linda Thienpont*[6] and *Callum G. Fraser*[7]

EQAS Organizers Working Group on Quality Goals

[1] Servicio de Bioquímica, Hospital General Vall d'Hebron, Barcelona, Spain
[2] Central Clinical Chemistry Lab, St. Radboud University Hospital, Nijmegen, The Netherlands
[3] Department Clinical Biology, Institut D'Hygiène et d'Epidémiologie, Brussels, Belgium
[4] Department of Clinical Chemistry, Odense University Hospital, Odense, Denmark
[5] Instand e. V., Düsseldorf, Germany
[6] Universiteit Gent, Faculteit Van de Farmaceutische Wetenschappen, Gent, Belgium
[7] Department of Biochemical Medicine, Ninewells Hospital & Medical School, Dundee, Scotland

Summary: A questionnaire was circulated to European countries seeking information on the criteria used for acceptable performance in external quality assessment schemes. Responses were obtained from 21 countries. Fixed limits are used in 13 countries but the basis for these varies widely and includes clinical decision making, biological variation, views of experts, the state-of-the-art, and combinations of approaches. Variable limits based upon statistical analysis of the performance attained are used in 8 countries. The many advantages of harmonization in Europe have prompted the development of criteria based upon within- and between-subject biological variation for use in schemes which circulate single specimen challenges. Currently used criteria, which show much diversity, are compared with these proposals, and the empirical nature of the majority of the former is demonstrated.

Introduction

This work was done by a Working Group which was formed in 1994, as consequence of various European EQAS-organizers' meetings held under the umbrella of the CEC Standards, Measurements and Testing Programme (formerly BCR). The aim of the Working Group is to collaborate in the harmonization of results in the field of laboratory medicine; it acts on a voluntary basis, in conjunction with three other Working Groups coordinated by Dr. *Adam Uldall*, of Herlev Hospital, Denmark.

At present, external quality assessment schemes (EQAS) exist in the field of laboratory medicine in many countries. Most of these are intended to assist individual laboratories to continuously monitor their performance, and to compare it with that of other laboratories, whereas others may be intended primarily for accreditation/licensing purposes. Additionally, EQAS may monitor the quality of commercial analytical systems, reagents and test kits, and they help manufacturers to achieve a better harmonization of the results from these different analytical techniques. Owing to these different aims and its different stages of development, the design of EQAS varies to a great extent in individual European countries. On the one hand this is an advantage, because it allows the EQAS to be adapted to the special prevailing situation in each country. On the other hand it creates difficulties with respect to the ongoing social and economic harmonization efforts within Europe and prevents the setting up of a uniform health care policy.

We stress that this paper is not related to the possible role of EQAS for accreditation purposes. Rather, it asserts that the main objective of EQAS is to help laboratories in the creation of quality and to promote transferability of results among European countries.

The two principal aims of EQAS are to define target values, and to define the limits for acceptance. The target values should be assigned from reference methods, but as only a few schemes follow these principles, target values are derived from the statistics of each survey (overall or method-group mean). Although the relevant differences among countries are pointed out in this paper, they do not constitute the basis of this work. The reason for this decision is that to promote changes on this subject implies that the design of programmes must be reviewed (management field), and the aim of the Working Group is primarily intended to stimulate thought (the inherent educational role of current EQAS).

In particular, it has been realized in the recent years that a major obstacle to the harmonization of European EQAS is the different acceptance of limits used in individual countries. These limits are the interval within which the results of an individual laboratory must lie to be considered as acceptable. In consequence, the Working Group tried to develop a concept for deriving acceptance limits for EQAS which should be generally applicable all over Europe. Before doing this, the group felt it of great importance to have a picture of currently used limits and to understand the reasons underlying their generation. For this purpose, a questionnaire was sent to European EQAS organizers, seeking information on the criteria for setting limits and the actual numbers used for general biochemical quantities. Data from 21 countries have been received and are presented here. In addition, the Working Group outlines a concept for deriving EQAS acceptance limits from biological variation, which is intended for use as a common European working basis for currently conducted schemes.

Results and Discussion

The information provided reveals two main types of criterion for defining EQA limits:

i) criteria based on biological variation, opinions of experts, "fixed" state-of-the-art, or combinations of these, leading to "fixed limits";

ii) statistical criteria applied to the outcome of each survey, leading to "variable limits" (real state-of-the-art limits) (tab. 1).

i) Fixed limits are used in 13 out of the 21 countries which responded to the questionnaire. But, as addressed above, the criteria for deriving them vary to a great extent:

Tab. 1 Criteria for defining limits in EQAS
Abbreviations:
CV_{bi} = within-subject coefficient of variation; CV_{wlab} = within-laboratory coefficient of variation; P_{95} = 95th percentile; clin = clinicians; CCV = chosen coefficient of variation

Country	Fixed limits	Country	Variable limits
Denmark	3 (1/2 CV_{bi})	Italy (Lombardia)	P_{95}
Netherlands	3 (1/2 CV_{bi})	Spain	P_{95}
Belgium	Biology	France	P_{95}, P_{99}
Germany	3 (CV_{wlab})	Portugal	P_{95}, P_{99}
Czech Republic	3 (CV_{wlab})	Iceland	Murex
Luxembourg	3 (CV_{wlab})	Greece	–
Finland (labquality)	Experts, P_{95}	Russia	–
Norway	Labquality	Sweden	–
Switzerland	Clin, analysts		
Croatia	2 (CV_{wlab})		
Lithuania	–		
Ireland	CCV		
United Kingdom	CCV		

– *Denmark* (DK) recommends three times half the within-subject biological variation (s_I), using the desirable analytical standard deviation goal (s_a) for routine methods, which is $s_a < 0.5\ s_i$ (1, 2). But, being well aware that many routine methods currently do not meet this desirable goal for analytical standard deviation (e. g. methods for sodium, chloride, calcium, protein), the resulting EQA limits are meant as targets to be reached in the future rather than for judgement of current performance.

– *The Netherlands* (NL) use principally the same approach as described above for Denmark. As in Denmark, some of the limits are used as an aim to be strictly applied only in the future. On the other hand, lower limits are used for some quantities when the respective methods perform much better than required by strict adherence to desirable analytical goals (e. g. lactate dehydrogenase, aspartate aminotransferase and alanine aminotransferase). A peculiarity of this EQA scheme is that, in contrast to the generalized approach of single specimen challenges, they conduct a multispecimen testing scheme with the establishment of the laboratory mean and standard deviation, computed from eight results obtained in each time period. Thereafter, the statistically expected percentage (equal to the score of the participant) of results residing in the range of target ± three times the tabulated within-subject biological variation, is calculated (3).

– *Belgium* (BE) also uses limits based on biological variation, while respecting desirable analytical goals according to the combined allowable bias and standard deviation limits as proposed by *Fraser & Hyltoft* (4). However, for quantities where current analytical performance does not meet these goals, the desirable EQA limits are substituted by practical ones, derived from the state-of-the-art, as proposed in the document produced by a working group of EGE-Lab (5).

– *Germany* (DE) uses limits which are three times the maximum within-laboratory standard deviation (s), which themselves were derived from the respective reference intervals (6); but, additionally, take into account the analytical state-of-the-art at the time when the German guidelines became mandatory (7). Unique in the German scheme is the use of reference method target values for many quantities which, in turn, sometimes necessitates higher EQA limits than in other countries.

– The *Czech Republic* and *Luxembourg* have adopted the German system.

– *Finland* (FI) (and also *Norway* which participates in the Finnish scheme) and *Switzerland* (CH) use acceptable limits set by experts, which take account of the

clinical decision, the biological variation and the 95% in limit analytical state-of-the-art. (It should be noted here that since January 1995 Denmark, Norway, Iceland and Finland have been using mutual limits).

– *Croatia* (CR) reported the use of limits which were twice the maximum within-laboratory CV, without explaining how the respective CV data were derived.

– *Lithuania* reported the use of fixed limits, but again without explaining the underlying concept.

– The *United Kingdom* (GB) uses average CV values based on historical observed data from the scheme, established around 20 years ago (CCV) (8), for participant assessment.

– *Ireland* (IE) has adopted the GB system, but classifies participants as poor only when their results are "far away" from those of the majority.

ii) eight countries base their limits for EQA on the actual outcome of each survey. Therefore, the values given in tables 2–4 represent an average of results from recently conducted surveys.

– *Spain* (ES), *Italy* (IT) (Lombardia), *France* (FR) and *Portugal* (PT) judge all results acceptable which fall within the 95% or 99% interval (depending on the quantity) around the mean.

– *Iceland* (IS) participates in a commercial scheme (Murex Diagnostics) which uses statistical acceptance criteria similar to those described above, but the actual limits were not reported.

– The EQA scheme in *Russia* is merely informative without using acceptance limits.

– *Sweden* started an EQA scheme as recently as 1992 and has not yet formulated acceptance limits; the same is true for *Greece*.

The limits reported by the different countries are presented in tables 2–4 (country grouping is identical to that in tab. 1). It should be noted here that most schemes work with single analysis of specimens and participant assessment in each survey (except the Netherlands and the United Kingdom, which use cumulative survey data for performance assessment). As mentioned above, no

Tab. 2 Currently used European EQA limits (given in % deviation from the target)

	Na	Cl	Ca	Mg	Albumin	Protein	Glucose	K	Creatinine
Denmark	0.9	2.1	2.7	3.5	4.2	4.2	6.6	8.2	6.6
Netherlands	0.9	2.1	2.7	3.3	4.2	4.2	10.0	7.2	6.6
Belgium	2.0	3.0	4.5	9.5	6.2	5.5	14.0	8.0	8.0
Germany[a]	6.0	6.0	10.0	12.0	18.0	9.0	15.0	8.0	18.0
Finland[a]	3.0	3.0	3.0	5.0	5.0	5.0	5.0	3.	5.0
Switzerland	2.0	3.0	4.0	4.0	6.0	3.0	7.0	3.0	15.0
Croatia	3.0	4.0	5.0	–	–	8.0	5.0	5.0	10.0
Lithuania	3.0	3.0	2.0	–	3.0	3.0	5.0	2.0	5.0
United Kingdom	1.6	2.2	4.0	10.0	7.5	3.9	7.7	2.9	8.9
Spain	6.6	10.0	10.0	–	14.0	9.2	9.8	7.4	14.0
Italy	2.0	4.0	5.5	–	4.0	4.0	6.0	3.0	8.8
France	3.5	4.0	4.6	12.0	10.0	10.0	11.0	6.8	11.0
Portugal	2.5	6.0	7.0	–	–	5.0	6.0	5.0	12.0

[a] same limits for Czech Republic and Luxembourg
[b] same limits for Norway

Tab. 3 Currently used European EQA limits (given in % deviation from the target)

	Cholesterol	P_i	Lithium	Lactate dehydrogenase	Urate	Alkaline phosphatase	Amylase
Denmark	8.1	12.0	–	12.0	13.0	10.0	11.0
Netherlands	8.1	–	5.0	3.0	10.0	8.0	10.0
Belgium	8.4	14.0	10.0	15.0	15.0	10.0	17.0
Germany[a]	18.0	15.0	12.0	21.0	18.0	21.0	21.0
Finland	5.0	5.0	5.0	10.0	5.0	10.0	10.0
Switzerland	3.0	10.0	6.0	15.0	10.0	15.0	20.0
Croatia	10.0	10.0	–	20.0	10.0	20.0	–
Lithuania	7.0	5.0	–	7.0	7.0	7.0	10.0
United Kingdom	7.6	7.8	11.0	13.0	7.7	15.0	11.0
Spain	9.8	12.0	22.0	17.0	15.0	22.0	56.0
Italy	5.5	9.5	–	10.0	8.0	18.0	–
France	16.5	–	10.0	20.0	16.0	20.0	25.0
Portugal	5.0	8.0	–	16.0	9.0	29.0	–

[a] same limits for Czech Republic and Luxembourg
[b] same limits for Norway

data are shown in tables 2–4 for the Czech Republic, Luxembourg, Norway and Ireland because they have adopted values from other countries. In addition, Russia, Sweden, Greece and Iceland are not represented, either because acceptance limits are not used in those countries, or because they were not reported. Further, quantities have been arranged according to increasing biological variation.

This principle was also used for creating figure 1, which is intended to give a rapid overview of the limits without indicating the countries applying them. In figure 1 also, the values derived from the concept of the Working Group (see below) are included, using the symbol "▲". As can be seen from figure 1 and tables 2–4, the currently used European EQA limits show relatively high variation for nearly all quantities. For example, for sodium they vary between 0.9% in Denmark and 6.6% in Spain (tab. 2), for cholesterol from 3% in Switzerland to 18% in Germany (tab. 2), and for urea from 5% in Finland to 24% in Germany (tab. 4). The same wide disagreement may be seen in figure 2, where data have been grouped according to the type of limits used: fixed limits on the left and variable limits on the right.

This is not surprising because the different EQA schemes have different aims and are conducted under different constraints. Countries basing their limits on biology (e. g. Denmark and The Netherlands) have narrow limits for analytes with a low biological variation and wide limits for analytes with high biological variation. But the former in particular are primarily intended as goals to be reached in the future. In practice, they are often widened for quantities with a narrow biological

Tab. 4 Currently used European EQA limits (given in % deviation from the target)

	Urea	Aspartate amino-transferase	Bilirubin	γ-Glutamyl-transferase	Triacyl-glycerol	Alanine amino-transferase	Fe	Creatine kinase
Denmark	19.0	22.0	34.0	22.0	34.0	41.0	48.0	62.0
Netherlands	19.0	7.0	33.0	18.0	33.0	10.0	30.0	63.0
Belgium	16.0	16.0	24.0	15.0	20.0	20.0	–	20.0
Germany	24.0	21.0	24.0	21.0	21.0	21.0	21.0	24.0
Finland[b]	5.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Switzerland	7.0	15.0	30.0	15.0	10.0	15.0	12.0	20.0
Croatia	7.0	20.0	10.0	20.0	10.0	20.0	10.0	20.0
Lithuania	7.0	7.0	–	10.0	7.0	7.0	5.0	7.0
United Kingdom	5.7	12.0	19.0	13.0	–	15.0	15.0	18.0
Spain	10.0	17.0	28.0	18.0	14.0	17.0	16.0	52.0
Italy	9.5	10.0	–	13.0	8.5	13.0	9.0	16.0
France	16.0	20.0	15.0	20.0	15.0	20.0	20.0	25.0
Portugal	6.0	12.0	13.0	11.0	7.0	11.0	7.0	14.0

[a] same limits for Czech Republic and Luxembourg
[b] same limits for Norway

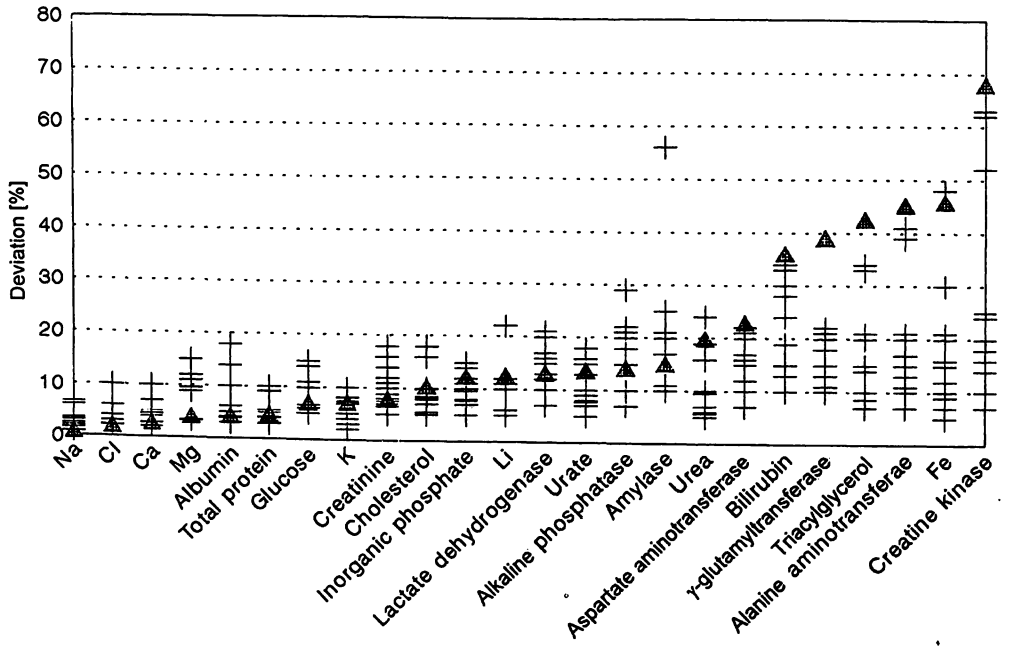

Fig. 1 Current European EQA limits.
Analytes arranged in ascending biological variation.

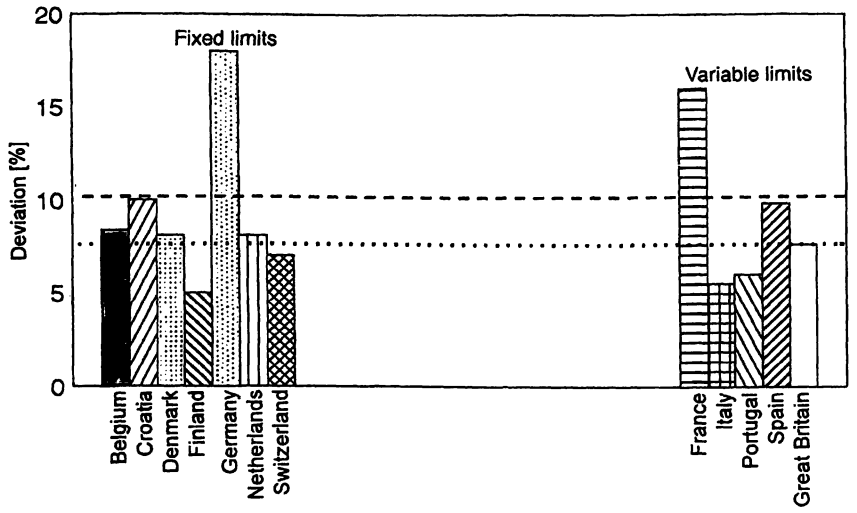

Fig. 2 Interlaboratory variation limits for cholesterol.

variation, in order to reach realistic acceptance figures which can be presented to the participants (Belgium and The Netherlands). Germany, with a mixture of biology and state-of-the-art limits, as well as providing a scheme acceptable by the health insurance system, has to use relatively large limits, except for quantities where the current analytical performance is much better than required by biology (e. g. enzyme activity assays). EQA limits set by experts (Finland and Switzerland) take account of current analytical performance, in turn leading to relatively wide limits for quantities with narrow biological variation like sodium, albumin or calcium. On the other hand, they show a tendency to set a general upper EQA limit which is 10% in the case of Finland. Interestingly, Lithuania follows the Finnish limits very closely, possibly because these two countries are geographically close. Croatia, did not reported the basis of its limits, and sets an upper limit of 20%.

Among the countries using variable limits, reflecting the "real" state-of-the-art, Spain generally shows the highest limits, while Italy mostly shows the lowest. This might be due to the different statistical levels applied for acceptance, the wide diversity of procedures used or the different types of laboratories participating (e. g. studies made in Spain revealed that in certain areas all laboratories use the same procedures with a consequent general agreement of results, and that the group of public laboratories had less variation than the overall group; in the case of Italy only one specific geographical area has submitted data to our questionnaire), or the different targets used (e. g. overall mean or group target). In addition, also in this group, there seems to be a tendency for setting upper limits (e. g. in France, 20–25% for enzymes). As pointed out above, the United Kingdom is unique because performance is judged from cumulated data, which mostly allows more narrow limits to be used than in the other countries in this group. Other issues such as common standardization (including calibrators, control materials with minimum matrix effect, etc) and reliable target values may also be addressed in this context. But these considerations are beyond the scope of this work.

The concept of the Working Group

Considering the data received, we believe that there is a strong need for harmonization of EQA limits in Europe. But, it is clear that harmonized European EQA limits are only possible with a harmonized analytical design of the schemes (e. g. single or multiple measurements, single target or multiple targets, 95 or 99% confidence interval).

Therefore, the Group first had to define the situation to which their concept should be applicable. Because most schemes use single measurements and certain cut-off values for judgement of performance, the Working Group restricted itself to this design. This does not mean that the Working Group recommends this approach for the future. On the contrary, it recommends development of alternative EQA models (9), more appropriate for instructive purposes, but which are out of the scope of this presentation. The model presented below, therefore, is primarily intended as a realistic working basis for EQA schemes as they are conducted today. In any case, the Working Group is convinced that a theoretical concept based on biology should be the starting point for deriving EQA limits for every situation. Moreover, EQA limits have to be built on quality specifications for routine methods. The Group therefore chose, as the principal underlying concept for deriving EQA limits, the desirable specifications for routine method bias and random error combined (5, 10), which are the sources of uncertainty affecting a single analysis. Then, the desirable EQA limits (or desirable maximum deviation of a participant from the target = D%) can be expressed as follows:

$$D < K \times 0.5\ CV_i + 0.25^{-}(CV_i^2 + CV_g^2)^{1/2}$$

K = 1.65 or 2.33 for 95 or 99% acceptance
CV_i = average within-subject biological variation
CV_g = average between-subject biological variation

We preferred the use of coefficient of variation over standard deviation because nearly all EQA schemes use the former.

According to this formula, the percentage deviations of a single analysis derived from biology (99% confidence interval) for the quantities studied are shown in table 5. Figure 1 shows that quantities with low biological variation (sodium, chloride, calcium and albumin) have narrow acceptance limits. At present very few countries maintain interlaboratory variation within these restricted intervals, but a general application of these limits would spur manufacturers to develop improved analytical procedures. However, we emphasize that other mechanisms may also be effective in attaining this laudable goal, e. g. introduction of the accuracy concept into test-kit production by comparison of results on patient's specimens with accepted accuracy-based reference methods (11).

Tab. 5 Percentage deviations of a single analysis derived from biology (99% confidence interval)

Quantity	Deviation (%)
Na	0.90
Cl	2.13
Ca	2.80
Mg	4.16
Albumin	4.36
Total protein	4.8
Glucose	7.0
K	7.2
Creatinine	7.9
Cholesterol	10.4
Inorganic phosphate	12.4
Li	12.6
Lactate dehydrogenase	13.2
Urate	13.8
Alkaline phosphatase	14.3
Amylase	15.1
Urea	20.8
Aspartate aminotransferase	23.1
γ-Glutamyl transferase	29.2
Bilirubin	36.2
Triacylglycerol	42.6
Alanine aminotransferase	45.3
Fe	45.9
Creatine kinase	68.0

The limits set for those quantities in which biological variation is intermediate (from protein to amylase in fig. 1) are attainable in most of the countries questioned, indicating that goals based on biology are realistic and feasible for use in EQAS, with the aim of harmonizing results.

Presently, the application of biology-derived limits is entirely practicable for the quantities with high biological variability, such as enzymes, urea, bilirubin, iron and triacylglycerols. However, if the current state of the art produces better precision and negligible bias, then additional benefits (such as reduction of costs in internal quality control) may be obtained if laboratories try to maintain such analytical quality. Another point is the appropriateness of limits derived from biology. For particular medical situations and clinical strategies, the analytical quality specifications might be different. These aspects have been discussed previously (12, 13). Such strategies, however, vary from country to country, so this group has taken the view presented by the EGE-lab group with minor modifications, as general analytical quality specifications for external quality assessment. When analytical quality specifications based on common clinical strategies are known, these should be used, but a detailed discussion of this item is outside the scope of this work.

Harmonization of quantities with narrower biological variation will involve much work and time. However, this Group concludes that it is an attainable objective and advocates the adoption of common goals based on biology: their application would allow common reference intervals and common reference changes (= medically significant differences) to be shared, with the consequent reduction in laboratory costs. It is evident that achieving concurrence in laboratory performance is an important step towards optimizing the utility of laboratory data throughout Europe.

References

1. Proceedings of the Subcommittee on Analytical Goals in Clinical Chemistry. World Association of Societies of Pathology, London. Analytical goals in clinical chemistry: their relationship to medical care. Am J Clin Pathol 1979; 72:624–30.
2. Harris EK. Statistical principles underlying analytic goal-setting in clinical chemistry. Am J Clin Pathol 1979; 72:374–82.
3. Steigstra H, Jansen RTP, Baadenhuijsen H. Combi Scheme: new combined internal/external quality-assessment scheme in The Netherlands. Clin Chem 1991; 37:1196–204.
4. Fraser CG, Hyltoft Petersen P. Quality goals in external quality assessment are best based on biology. Scan J Clin Lab Invest 1993; 53 Suppl 212:8–9.
5. Fraser CG, Hyltoft Petersen P, Ricós C, Haeckel R. Proposed quality specifications for the imprecision and inaccuracy of analytical systems for clinical chemistry. Eur J Clin Chem Clin Biochem 1992; 30:311–7.
6. Stamm D. A new concept for quality control of clinical laboratory investigation in the light of clinical requirements and based on reference methods. J Clin Chem Clin Biochem 1982; 20:817–24.
7. Richtlinien der Bundesärtzekammer zur Qualitätssicherung in medizinischen Laboratorien. Deutsches Ärzteblatt 1988; 85:A699–A712.
8. Bullock DG. External quality assessment in clinical chemistry: an examination of requirements, applications and benefits. Thesis submitted to the Faculty of Science of the University of Birmingham for the Degree of Doctor of Philosophy, Chapter 15:371–84, 1987.

9. Libeer JC, Baadenhuijsen H, Fraser CG, Hyltoft Petersen P, Ricós C, Stöckl D, Thienpont L. Characterize and classify various types of EQA according to the objectives. Draft paper, European EQA-organizers Working Group A: "Analytical goals in laboratory medicine".
10. Stöckl D, Baadenhuijsen H, Fraser CG, Libeer JC, Hyltoft Petersen P, Ricós C, Thienpont L. Desirable routine analytical goals for quantities assayed in serum. Discussion paper from the members of the External Quality Assessment (EQA) Working Group A[1]) on analytical goals in laboratory medicine. Eur J Clin Chem Clin Biochem 1995; 33:157–69.
11. Thienpont L, Franzini C, Kratochvila I, Middle J, Ricós C, Siekmann L, et al.: (Editorial Board WG-B). Analytical quality specifications for reference methods and operating specifications for networks of reference laboratories. Draft paper, European EQA-Organizers Working Group B: "Target values in EQAS".
12. De Verdier CH, Groth T, Hyltoft Petersen P. Medical need for quality specifications in clinical laboratories. Uppsala J Med Sci 1993; 98:187–491.
13. Hørder H, editor. Assessing quality requirements in clinical chemistry. The Nordic Clinical Chemistry Project NORD-KEM). Helsiniki, Finland 1980:144 pages.

Received April 10/October 18, 1995

Corresponding author: Dra. Carmen Ricós, Servicio de Bioquímica, Hospital General Vall d'Hebron, P. Vall d'Hebron, 119–129, E-08035 Barcelona, Spain